COURSERA FINAL PROJECT

SPECIALIZED MODELS: TIME SERIES AND SURVIVAL ANALYSIS

Time Series Forecasting to predict grocery sales at Favorita stores

Project presented by

Raziel Amador Rios

to obtain the Coursera Certificate

Main object: produce a reliable forecasting based on sales time-series data from a retail store (Favorita stores, an Ecuadorian-based company).

Barcelona, January 2023

ii

Table of Contents

iv

Time Series Forecasting

I

Main objective

The **main object** of the project is to generate a reliable **time series forecasting** based on sales from Favorita stores, which is a grocery retail store based in Ecuador. Producing an accurate forecast could lead to decreased food waste related to overstocking and improve customer satisfaction.

All the code can be found in the following GitHub repository, further information can be found in Supplementary information.

II

Data description

The dataset comes from **Kaggle**, named: Store Sales - Time Series Forecasting. The Kaggle dataset contains sales data from *Corporación Favorita*, a large Ecuadorian-based grocery retailer. Kaggle provides you with 7 diffent files (see Table 1 for more details).

Number	File Name	Description
1	holiday_events.csv	Relevant holidays in Ecuador
2	oil.csv	Oil prices from 2013 to 2017
3	sample_submission.csv	submission example
4	stores.csv	Stores metadata
5	test.csv	Stores and family products
6	train.csv	Sales by store and product-family from 2013-01 to 2017-08
7	transactions.csv	Number of transactions by store

Table 1: Kaggle files description. Files are alphabetically sorted.

The training data represents 99% of the data, including dates from 2013-01-01 to 2017-08-16 (55.5 months), 54 stores placed in different cities within Ecuador, and 33 family-products (see Figure 1). The testing data includes dates from 2017-08-16 to 2017-08-31 (15 days).

Stores Summary					
54	**5**	**17**	**33**	**16**	**55.5**
Stores	Store types	Store clusters	Product families	States	Months

Figure 1: Summary of the training dataset. The cluster information denotes similarity between stores.

III

Data exploration and data cleaning

Appendix

IV

Supplementary information

This work was written with emacs[1] using LaTeX[2], using only **Free and Open Source software**. All the computational analysis were carried out using Linux-based distributions. The figures were generated with Python (matplotlib[3]/seaborn[4]/plotly[5]) and Inkscape[6].

Contact: razielar@gmail.com

---

[1]https://www.gnu.org/software/emacs/
[2]https://www.latex-project.org/
[3]https://matplotlib.org/
[4]https://seaborn.pydata.org/
[5]https://plotly.com/
[6]https://inkscape.org/